標準コード化による傷害事例テキストからのベイジアンネットモデル構築
— ICD・ICF・JICFS コード変換モデルを用いた傷害危険性予測システム —
Bayesian Network construction from normalized injury data
Prediction of potential injury hazards by using ICD,ICF,JICFS normalization code

池田 涼太郎[*1] Ryotaro Ikeda　三浦 未生[*3] Miki Miura　本村 陽一[*2] Yoichi Motomura　西田 佳史[*2] Yoshifumi Nishida　原 一之[*1] Kazuyuki Hara

[*1]東京都立産業技術高等専門学校 Tokyo Metropolitan College of Industrial Technology
[*2]独立行政法人 産業技術総合研究所 National Institute of Advanced Industrial and Technology
[*3]産業技術大学院大学 Advanced Institute Of Industrial Technology

Unintentional injuries have been first leading cause of children 's death in Japan. This situation has been constant for scores of years. One of the causes that this situation hasn 't changed is that information related to injuries is not shared and doesn 't become reusable knowledge. To solve these problems, a method for extracting reusable knowledge from injury data and sharing in society is necessary. But since injury data are not described in a standardized format, applying statistical approach to the data directly for analysis is difficult. The authors took an approach that we defined standard codes and statistical methods can be applied by converting the injury data with created codes. In this paper, we proposed a method which enables statistical methods can be applied to children 's injury data. We also constructed Bayesian network model using converted data with the codes and reported some examples of analysis using the constructed model.

1. 緒論

1 歳から 19 歳までの日本の子供の死亡原因として最も多いのは不慮の事故である．日本ではこれらの事故の経験を活かし, 事故予防に役立てるための取り組みが十分に確立されていないためである. また, 海外では事故情報の収集は行われているものの, 事故の割合の算出などの単純な統計処理のみで, 事故予防にまで十分に活用されていない. そのため, こうした事故情報が十分に共有されず, 過去に同じような原因で起きた事故が度々繰り返されており, 大きな問題となっている.

事故情報を事故予防に活かすためには, 統計的手法を利用し, 再利用可能な形で, 知識化し, 共有していく必要がある. しかし, 事故情報を共有し統計的手法に利用するには, 項目の一つ一つが離散変数化され, 事象として数えられる形式になっている必要がある. このため, その際に項目の事象を離散変数化する為に特定の分類規則が必要となるが, 現状では蓄積された多くの統計情報は分類処理がなされていないか, 独自の分類規則に従って分類されている為共有化を行う事は難しい. そこで筆者らは, 分類規則を統一する為に標準コードと呼ばれる分類規則を既存の統計情報に適応することを考えた. 標準コードとは, 国際的な, あるいはそれに順ずる程広い範囲で使われる事を想定した分類規則である.

本研究では, 国立成育医療センターで収集された事故事例の情報を標準コードを用いて, 統一化された表現に変換し, それを用いてベイジアンネットワークモデルを構築して事故予防に有益な情報を提供する手法を提案する.

連絡先: 池田 涼太郎, (独) 産業技術総合研究所 デジタルヒューマン研究センター〒 135-0064 東京都江東区青海 2-41-6, 電話 : 03-3599-8001 , FAX : 03-5530-2061, r-ikeda@aist.go.jp

2. 事故事例データの標準コード化

2.1 事故事例データの概要

国立成育医療センターより提供された約 3000 件の事故事例データには, 子供の年齢や性別等に加え, 診察の際に患者から聴取した事故へ至るまでのいきさつ (以降自由記述データと呼ぶ) が書かれている. 自由記述データには, 具体的には (1) どのような行動を取っていて (2) どのような物が原因で (3) どのような事故に至ったか 等の情報が含まれている. 事故事例データについての詳細は [本村 06] を参照のこと.

2.2 事故事例データの統計分析における問題点

自由記述データは, 普段我々が文章を読み書きする際に用いる自然言語によって記述されている. 具体的には次のような文章である.

> 母は不在だった．23：00 帰宅して歯磨きをしていて口腔内の出血，創に気づいた．
> 留守番をしていた父を問いただしたところ，転倒，打撲を知ることになった．

このような情報は, そのままでは統計処理を適応することが出来ない為, 文章中から, 怪我の種類, 原因となった物, 行動等の要点を抽出し, それらを離散変数として表現しなくてはならない. 上記の例でいうと, 怪我の種類が口腔内の出血と打撲, 行動が歯磨きとなる. ここで, 文章表現の違いによって生じうる表記揺れも考慮しなくてはならない. 例えば,「口腔内の出血」が「口から血が出た」と記述されていたり,「歯磨き」が「歯を磨いていた」と記述されている可能性もある. この情報を離散変数化し, 更に情報共有を可能にするためには, 統一化された標準フォーマットで表現可能にする必要がある. このため, 統計的手法を適応するには自然言語から要点となる言葉を抜き出す為に, 表記揺れに対応し, それを標準コードに変換する作業が必要となる.

2.3 事故事例データの標準コード化

より広い範囲での情報共有を可能にする為, 標準コードは様々な分野の範囲で用いることができ, かつ分類分けにあいまいさが生じたり, 分類の際に漏れが生じたりしないものである必要がある. 本稿では標準化コードとして, 疾病を分類するための国際分類コードである ICD (参考:[疾病, 傷害及び死因分類]), 人間の生活機能を分類するための国際分類コードである ICF (参考:[ICF 及び ICF-CY の活用]), 流通業界で商品の分類コードとして使われている JICFS (参考:[データベースサービス: JICFS/IFDB]) を用いた.ICD,ICF は世界保健機関 (WHO) によって制定された国際規格であり,JICFS は JIS 規格になっている為信頼性が高く, 標準コードとして利用するのに適している.

これらの標準コードは全て, それぞれの項目が階層構造になっている. 階層が浅くなれば広い意味を, 階層が深くなれば詳細な意味を表すコードになっており, 意味の範囲をある程度明示できるのが特徴である. 例えば,「遊ぶ」という言葉を「レクリエーション」という意味でのみ解釈するか, 野内外問わず友達と遊んだりする意味まで含めるか, 等を明示することができる. これは, 事故事例データにおける表現がある程度抽象的であってもコード化に対応できるようにする為に重要である.

自由記述データを標準コードに変換する為に, 検索エンジン等で表記揺れに対応する為に利用されている, 辞書を使って文章中の統一化されていない表現の言葉を標準コードに変換する手法を用いた. 辞書とは, 特定の単語とそれの類義語をリストアップした対応表のことである. 具体的な辞書の例として,JICFS コードへの変換辞書の一部を表 1 に示す. この表において「嗜好飲料」と「トレーニング器具」が JICFS コードであり, 二重線より右側の「ジュース」や「お茶」,「鉄アレイ」や「ベンチプレス」等が変換対象の言葉である.

表 1: JICFS コードへの変換辞書の一部

嗜好飲料		ジュース	お茶
トレーニング器具		鉄アレイ	ベンチプレス

標準コードへの変換には, この対応表をそれぞれの標準コード用に用意して用いる. 具体的には, 自由記述データから辞書における「類義語」を探索し, 対応する単語を標準コードとして関連付ける. この手法を用いることで, 文章の要点となる単語を抽出し, それを標準コードに対応させることが可能となる.

辞書は次の手順で作成した. まず, オープンソース形態素解析エンジン [MeCab] を用いて約 3000 件の自由記述データに対して形態素解析を行い, 文章に含まれる必要な単語を抽出する. この作業は, 自動化を行うためのソフトを作成し, それを利用した. 次に, それらに対応する標準コードを見つけ, それらの対応を辞書として記述する.

3. 標準コードを用いた傷害モデリング

事故データを標準化コードを用いて統一化された表現に変換したデータを用意し, それを用いてベイジアンネットワークモデルの構築を行った. モデルの構築には, ベイジアンネットワーク構築支援システムである Bayonet を用いた.

3.1 モデリングに用いた変数

確率変数として, 子供の行動指向に影響しているであろう”年齢”(0-3 歳,4-6 歳,..),”性別”(男, 女), それぞれ標準コード化した”事前行動”(走ること, 移動,..),”物”(嗜好飲料, ファンヒーター,..),”怪我”(挫創, 熱傷,..) の 5 つを用いた.

3.2 標準コードを用いた傷害危険性のモデル化

5 つの項目中, 年齢, 性別は生データをそのまま使った. 残りの事前行動, 物, 怪我については標準コードに変換したものを用いた.

1 つのノードの状態数が多数になると, それぞれの状態におけるリンクの関係が 1 つに纏まってしまう為, 全ての変数の状態を True/False の二項の事象の組み合わせで表せるように変換 (クラスタリング) した. 例えば, 性別であれば「男/女」ではなく,「男である/男でない」,「女である/女でない」のように表す. クラスタリングを行ったデータを用いてモデルを構築することで,1 つのノードが 1 つの状態に対応することになり, 状態同士の確率的因果関係をより正確に見る事が可能となる. 上記のようにクラスタリングを行ったデータを用いてモデル構築を行った結果を, 図 1 に示す. 尚, 評価基準には AIC(赤池情報基準) を用いた.

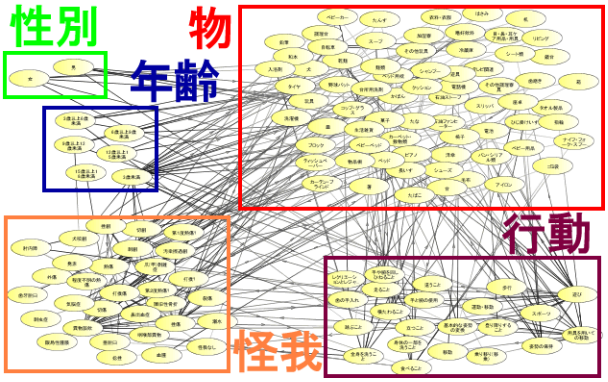

図 1: 構築したベイジアンネットワークモデル

構築したベイジアンネットワークモデルのリンクの引かれ方から分かることとして, 物と怪我のノードに「3 歳未満」のノードからのリンクが圧倒的に多く引かれており, 他の年代と比べて圧倒的に事故の危険性, 並びに「物」への関連性が強いことなどが挙げられる.

3.3 傷害危険性モデルによる推論の考察

行動や物を表すノードに状態値をセットし, 確率推論を実行させる. その結果, 怪我を表すノードそれぞれの事後確率がどのように変化するかを観測することで, 子供の行動や, 子供の近くにある物がどのような事故とどの程度の確率的な依存関係を持っているのかを定量的に調べる.

このような操作を行う事で, 例えば次のような事が分かる.

- 行動:スポーツ中=True , 怪我:挫傷, 打撲傷等の傷病=True とした場合, 9 歳から 15 歳の間の子供が特に怪我をしやすい傾向が見られた. この推論結果を図 2 に示す。

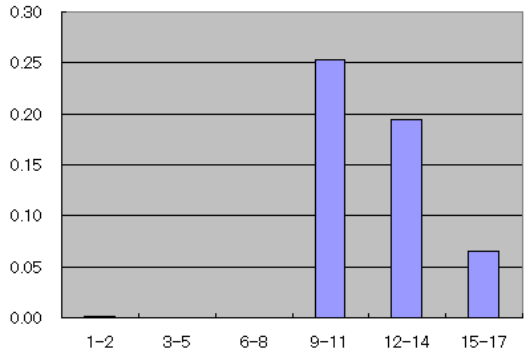

図 2: スポーツ中における各年齢別にみた怪我のしやすさ

※縦軸:怪我のしやすさ (事後確率) , 横軸:年齢層

- 行動を表すノードに 1 つずつ True の値をセットしていくと, 子供が火傷をする可能性は, 食事が他の行動に比べて高い傾向が見られた. この推論結果を図 3 に示す。

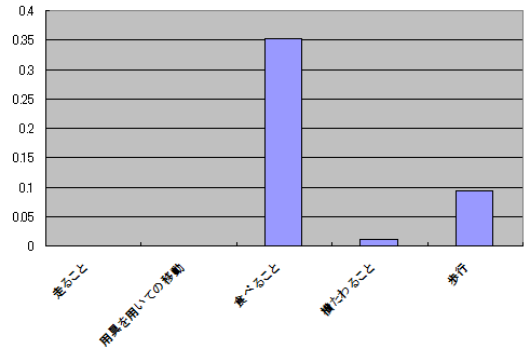

図 3: 行動別にみた子供の火傷のしやすさ

※縦軸:火傷のしやすさ (事後確率) , 横軸:行動

- 物:皿=True , 行動を表すノードそれぞれに 1 つずつ True の値をセットしていくと, 切創事故が食事中に起こりやすくなる傾向が見られた. この推論結果を図 4 に示す。
※縦軸:切創事故の起こりやすさ (事後確率) , 横軸:行動

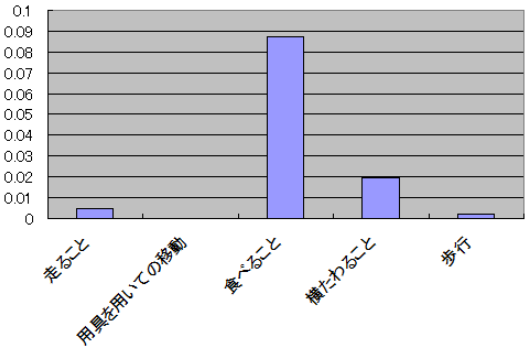

図 4: 行動別にみた、皿による切創事故の起こりやすさ

4. 結論

本稿では, 自由記述データに統計的手法を適応する為に自由記述データから文章の要点となる単語を抜き出し, それらを標準コードを用いて, 統一された表現に変換し, それを用いてベイジアンネットワークモデルを構築する手法を紹介した. その手法は次の通りである:(1) 自由記述データに対して形態素解析を行い, 必要な単語を抽出する (2) それらの単語と標準コードを対応させる為の辞書を作る (3) 辞書を用いて, 自由記述データを標準コードに関連付ける (4) クラスタリングを行い, 確率的モデルを構築する.

また, 構築したモデルを実際に利用して, 傷害データから事故予防につながる有益な情報を得た. この方法を用いて, 行動や物が各種の事故に影響する定量的な確率的因果関係を標準コード化した再利用性の高いベイジアンネットにより計算し, 潜在的な傷害危険性を評価することができた. 今後, 今回作成したモデルを用いた潜在的な危険性の予測を誰にでも行えるように,Web 上で動作するシステムの開発を行う.

参考文献

[本村 06] 本村, 西田, 北村, 金子, 柴田, 溝口 : 知識循環型事故サーベイランスシステム, 統計数理, (2006) .

[Miura08] 三浦, 本村, 柴田, 西田, 山本 : 事故サーベイランスシステムからの知識獲得－テキスト情報からの確率的因果構造のモデル化－ , 人工知能学会,(2008)

[Bayonet] (独) 産業技術総合研究所 ,(株) 数理システム : Bayonet ,(http://www.msi.co.jp/BAYONET/), (2009)

[MeCab] 京都大学情報学研究科 , 日本電信電話株式会社コミュニケーション科学基礎研究所共同研究ユニットプロジェクト : オープンソース形態素解析エンジン MeCab , (http://mecab.sourceforge.net/),(2009)

[疾病, 傷害及び死因分類] 厚生労働省 : 疾病, 傷害及び死因分類 ,(http://www.mhlw.go.jp/toukei/sippei/index.html), (2009)

[データベースサービス: JICFS/IFDB] 財団法人 流通システム開発センター : データベースサービス:JICFS/IFDB , (http://www.dsri.jp/company/jicfsifdb/top.htm),(2009)

[ICF 及び ICF-CY の活用] 国立特別支援教育総合研究所 : ICF 及び ICF-CY の活用 , 国立特別支援教育総合研究所,(2005)